Least-Squares Regression Line

The LSRL is a model used to represent a set of quantitative data. Suppose you find the distance from each point in the data to the linear model, then square those distances and find the sum. This is called the sum of the squares of the residuals The Least-Squares Regression Line (LSRL) is the line that minimizes this sum. The equation of the LSRL is $\hat{y} = b_0 + b_1 x$.

x represents explanatory variable (actual data).

$\hat{y}$ represents predicted y-value.

b_0 represents y-int.

b_1 represents slope.

Given a set of data, you can calculate the LSRL (without using your calculator!). Knowing the correlation makes this task even easier. Use the following formulas:

$$b_1 = r\left(\frac{s_y}{s_x}\right) \qquad r = \frac{\sum z_x z_y}{n-1} \qquad s_x = \sqrt{\frac{\sum (x-\bar{x})^2}{n-1}} \qquad b_0 = \bar{y} - b_1\bar{x}$$

Let's try one!... The following ordered pairs represent the scores for five former statistics students on the Unit I test and the semester exam, respectively: (76, 82), (78, 75), (84, 88), (65, 71), and (79, 85). Calculate the LSRL (without using your calculator) for predicting semester exam grades.

L1	L2	L3	L4	L5		L6		
x	y	z_x	z_y	$z_x z_y$	$(x-\bar{x})$	$(x-\bar{x})^2$	$(y-\bar{y})$	$(y-\bar{y})^2$
76	82	−.057	.255	−.0145	−1.4	1.96	1.8	3.24
78	75	0.228	−.738	−.1681	1.6	2.56	−5.2	27.04
84	88	1.082	1.106	1.1976	7.6	57.76	7.8	60.84
65	71	−1.624	−1.305	2.1188	−11.4	129.96	−9.2	84.64
79	85	.370	.681	.2521	2.6	6.76	4.8	23.04
				sum: 3.386		sum: 197.2		sum: 198.8

r = 0.846

s_x = 7.0214

s_y = 7.0598

b_1 = 0.8499

b_0 = 15.336

define x: score on Unit I test

define y: score on final exam

LSRL: $\hat{y} = 15.336 + 0.849x$ $\quad \hat{final} = 15.336 + 0.849\ unitI$

(Now verify your results using the linear regression option on your calculator).

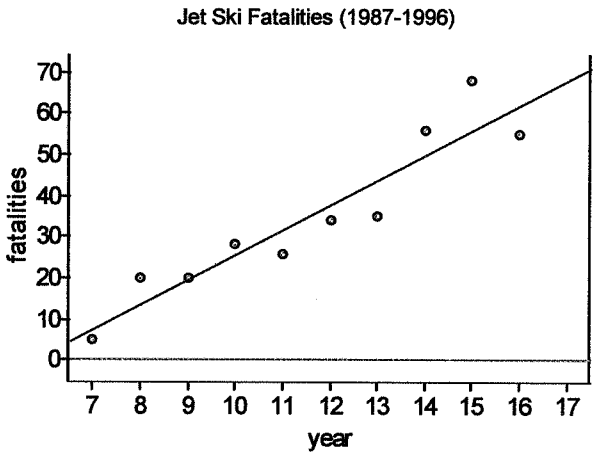
Coefficient of Determination

The coefficient of determination, also called R^2, is the square of the r-value (correlation). The R^2 value tells how much of the variation in the response variable is accounted for by the linear regression model. For example, if $R^2 = 1$, then 100% of the variability in the response variable is accounted for by the linear model. In other words, the relationship between the two variables is perfectly linear. If $R^2 = 0.95$, we can conclude that 95 % of the variability in the response variable is accounted for by the linear relationship with the explanatory variable.

7. Given the following set of data, find the equations of the LSRL, then find and interpret both the correlation and the coefficient of determination.

Jet Ski Use

	year	fatalities
1	7	5
2	8	20
3	9	20
4	10	28
5	11	26
6	12	34
7	13	35
8	14	56
9	15	68
10	16	55



a. LSRL: $\hat{fatal} = -34.648 + 6.03\,year$ (use meaningful variables in your equation rather than x and y, and use proper statistical notation!)

b. Correlation (*r-value*): 0.938. A correlation of 0.938 indicates that there is a strong, positive, linear relationship between year and number of fatalities.

c. Coefficient of determination (R^2): 0.880. An R^2 value of 0.880 indicates that 88 % of the variability in number of fatalities is accounted for by the linear relationship with the year.

8. A study of class attendance and grades earned among first-year students at a state university showed that in general students who attended a higher percent of their classes earned higher grades. Class attendance explained 16% of the variation in grades among the students. What is the numerical correlation between percent of classes attended and grades earned? 0.4

* check if $r > 0$ or $r < 0$

Residual Plots

A residual is the difference between the observed y-value and the predicted y-value for a given x-value.

residual = $y - \hat{y}$

The sum of squares of residuals (SSR) is used to determine the Least-Squares Regression Line for a given set of data.

A residual plot is a scatterplot which graphs the residuals on the vertical axis and the values of the explanatory variable on the horizontal axis for each data point, $(x_i, y_i - \hat{y}_i)$.

The residual plot gives a visual representation of the amount of error in the model. The closer the residuals are to zero, the smaller the error and the more accurate the model.

The LSRL is a good model if the residual plot shows random scatter relatively close to the horizontal axis (zero). The horizontal axis represents the LSRL.

Points in the residual plot that lie directly on the horizontal axis lie directly on the LSRL.

Points in the residual plot that lie above the horizontal axis lie above the LSRL. Therefore, the model gives an underestimate at that point.

Points in the residual plot that lie below the horizontal axis, lie below the LSRL. Therefore the model gives an overestimate at that point.

The LSRL is not a good model if the residual plot shows a pattern.

9. Construct a well-labeled residual plot using the data on jet ski fatalities from #7. What can you conclude about the appropriateness of the linear model based on the residual plot?